# Outlying Observation Diagnostics in Growth Curve Modeling

Xin Tong
University of Virginia

Zhiyong Zhang
University of Notre Dame

Tong, X., & Zhang, Z. (2017). Outlying Observation Diagnostics in Growth Curve Modeling. Multivariate Behavioral Research, 52(6), 768–788.

The research is supported by through the grant program on Statistical and Research Methodology in Education from the Institute of Education Sciences of U.S. Department of Education (R305D140037).

## Abstract

Growth curve models are widely used for investigating growth and change phenomena. Many studies in social and behavioral sciences have demonstrated that data without any outlying observation are rather an exception, especially for data collected longitudinally. Ignoring the existence of outlying observations may lead to inaccurate or even incorrect statistical inferences. Therefore, it is crucial to identify outlying observations in growth curve modeling. This study comparatively evaluate six methods in outlying observation diagnostics through a Monte Carlo simulation study on a linear growth curve model, by varying factors of sample size, number of measurement occasions, as well as proportion, geometry, and type of outlying observations. It is suggested that the greatest chance of success in detecting outlying observations comes from use of multiple methods, comparing their results and making a decision based on research purposes. A real data analysis example is also provided to illustrate the application of the six outlying observation diagnostic methods.

# Outlying Observation Diagnostics in Growth Curve Modeling

Growth curve (GC) models, as one of the fundamental tools for dealing with longitudinal data as well as repeated measures, are frequently used for investigating growth and change phenomena in social, behavioral, and educational sciences (e.g., McArdle and Nesselroade, 2014; Zhang et al., 2012). GC modeling allows examinations of intraindividual change over time as well as interindividual variability in intraindividual change. It is appealing not only because of its ability to model change but also because it allows investigation into the antecedents and consequents of change. Among methods developed for GC modeling, the normal-distribution-based maximum likelihood (NML) is routinely used and is incorporated in almost all statistical software. When a sample come from a normal population, NML generates consistent and efficient parameter estimates. However, practical data usually violate the normal assumption. For example, Micceri (1989) investigated 440 large scale data sets in psychology and found that almost all of them were significantly nonnormal. The occurrence of outlying observations in GC modeling is naturally more common because of the involvement of longitudinal data. When data are contaminated or contain outlying observations, NML estimates can be very inefficient or even biased (Yuan and Bentler, 2001), and Heywood cases or improper solutions may be created (Bollen, 1987).

Strategies to handle outlying observations have been developed. First, since outlying observations cause a problem especially encountered in models based on a limited number of individuals, a straightforward strategy is to observe more individuals in the population of interest. With more data collected, the underlying distribution of the sample can be better described, and it may turn out that we observe several additional data with extreme values so that the original outlying observation is no longer an outlying observation. Second, besides collecting more individuals, obtaining additional measurements for each individual may also account for the outlying observations, because the presence of multivariate outlying observations may indicate one or more important variables were omitted from the model (Lieberman, 2005). Third, human error often occurs in collecting data or processing the raw data, such as errors in entry, coding,

and transcription, and these errors may lead to extreme scores on one or more variables in the dataset. Thus, checking data consistency might be a solution to deal with outlying observations. The fourth strategy is to improve the model specification. If the data are used to estimate too complex models, or if the parameterization is incorrect, outlying observations are more likely to have larger effects. The fifth strategy is to conduct data transformation or directly remove outlying observations before data analysis (see Osborne and Overbay, 2004 for a more thorough discussion). Sixth, instead of direct transformation or truncation, researchers have developed various robust procedures to protect their data from being distorted by the presence of outlying observations. These methods either downweight the potential outlying observations as a transformation technique (e.g., Yuan and Bentler, 2000; Yuan and Zhang, 2012a) or assume that the data come from certain nonnormal distributions such as $t$ distribution or a mixture of normal distribution (e.g., Muthén and Shedden, 1999; Tong and Zhang, 2012). Among these strategies, the first four cannot be generally and easily applied. It is not always feasible to collect more data, obtain additional measurements, return to raw data to check consistency, or adapt model complexity and change parameterization. In practice, researchers usually transform the data so that they are close to being normally distributed or simply delete outlying observations prior to fitting a model to their datasets. Recently, more and more researchers (e.g., Savalei and Falk, 2014; Yuan and Zhang, 2012a) recommended the application of robust methods and statistics. Regardless of the strategy used, it is crucial to identify outlying observations in a dataset in the first place in order to obtain a better model estimation or interpret the extreme scores. Note that two methodologies with varying purposes are related to outlying observation detection. One is sensitivity analysis where data are assumed to be correct and we calibrate the model accordingly. In contrast, we may assume that the model is correct. If the person fit is not good, the corresponding case is identified as an outlying observation. This article aligns with the second methodology. In psychology, confirmatory data analyses are often conducted and a model is built based on a substantive theory. So we believe the model to be correct or at least useful, but data can be problematic. We are interested in detecting observations that are most unlikely to occur

under the hypothesized model. The outlying values in the data may lead to biased parameter estimates for the model and misleading model fit indices and test statistics.

The importance of outlying observation detection in multivariate data analysis has been recognized and various studies for detecting multivariate outlying observations have been conducted (e.g. Becker and Gather, 2001; Filzmoser et al., 2005; Peña and Prieto, 2001; Rocke and Woodruff, 1996; Rousseeuw and van Zomeren, 1990). A commonly applied method in those studies is to calculate a distance (i.e., Mahalanobis distance) from each point to the "center" of the data. An outlying observation is a point with a distance larger than some predetermined cutoff. For GC modeling, outlying observation detection is even more important because not only it can help improve the accuracy and precision of the model estimation, but also the detection procedure itself is very meaningful. It may help identify individuals who behave differently from the majority of the cases in a longitudinal study. Furthermore, it can tell whether an individual's growth pattern is different from the overall pattern and whether this individual only has extreme scores at some time points, e.g., talented students in the long run, or cheaters in a single test. Despite the increasing popularity of GC models and the fast growing interest in multivariate outlying observation detection, diagnostic tools to detect outlying observation in GC modeling lag behind. As far as we are aware, only Pan and Fang (2002) have specifically discussed how to detect outlying observations in the GC modeling framework. Although McArdle (1998) pointed out that an individual-level structural equation modeling approach permits a thorough analysis of outliers or subgroups, no systematical analysis has been conducted.

Because GC models can be fitted under the structural equation modeling framework (Meredith and Tisak, 1990), model diagnostic methods in structural equation modeling can be applied. In the framework of structural equation modeling, Bollen and Arminger (1991) developed a procedure using case-level residuals to identify outliers. Cadigan (1995) and Lee and Wang (1996) identified the most influential cases for the likelihood ratio statistics by applying the local perturbation procedure of Cook (1986) to structural equation modeling. The EQS software (Bentler, 1995) identifies cases that contribute most to Mardia's measure of multivariate kurtosis

and allows users to delete cases from analysis. To avoid the so-called masking effect where an outlying observations exists but is not identified or multiple outlying observations exist but not all of them are identified, Yuan and Zhong (2008) formally defined leverage observations and outliers in factor analysis and showed that robust procedures overcome the masking effect associated with procedures based on sample moments. Yuan and Hayashi (2010) then introduced two scatter plots for model diagnosis in structural equation modeling and Yuan and Zhang (2012b) further developed an R package `semdiag` to easily draw the two plots.

Based on the previous literature, we investigate six representative methods for multivariate outlying observation detection in GC modeling in this article. A univariate detection tool is first applied as a baseline for comparison. A traditional multivariate outlying observation diagnostic tool based on Mahalanobis distance and the method in Pan and Fang (2002) are applied to GC models as well. Then, we propose and apply three methods to study individual-level residuals and latent growth coefficients to not only identify outlying observations, but also distinguish two different types of outlying observations: leverage observations and outliers. We aim to evaluate and compare the performance of the six methods under different conditions. As far as we know, no study has systematically investigated and compared outlying observation diagnostic methods in GC modeling or multilevel modeling, let alone distinguishing leverage observations and outliers in that framework. To make this article self-contained, in the next section, we introduce the definition of two different types of outlying observation in GC models. The distinction between outlying observations and influential observations is highlighted. The subsequent section discusses the six methods that we use to detect multivariate outlying observations. Then, focusing on a linear GC model, a Monte Carlo simulation study is implemented to evaluate the performance of those methods. An example is also provided to illustrate the application of them, using data on the Peabody Individual Achievement Test (PIAT) mathematics assessment from the National Longitudinal Survey of Youth 1997 Cohort (Bureau of Labor Statistics, U.S. Department of Labor, 2005). We conclude the article by discussing the merit of each method and providing recommendations.

## Outlying Observations in GC Modeling

A GC model represents repeated measures of dependent variables as a function of time. In GC modeling, the relative standing of an individual at each time is modeled as a function of an underlying growth process, with random coefficients (e.g., initial level and rate of change) for that growth process being fitted to each individual. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ be a $T \times 1$ random vector and $y_{ij}$ be an observation for individual $i$ at time $j$ ($i = 1, \ldots, N; j = 1, \ldots, T$), where $N$ is the sample size and $T$ is the total number of measurement occasions. A typical form of unconditional GC models can be expressed as

$$\mathbf{y}_i = \mathbf{\Lambda}\mathbf{b}_i + \mathbf{e}_i, \quad (1)$$

$$\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{u}_i, \quad (2)$$

where $\mathbf{\Lambda}$ is a $T \times q$ factor loading matrix determining the growth trajectories, $\mathbf{b}_i$ is a $q \times 1$ vector of random effects, and $\mathbf{e}_i$ is a vector of intraindividual measurement errors. The vector of random effects $\mathbf{b}_i$ varies for each individual, and its mean, $\boldsymbol{\beta}$, represents the fixed effects. The residual vector $\mathbf{u}_i$ represents the random component of $\mathbf{b}_i$. In traditional GC analysis, it is assumed that the random effects $\mathbf{u}_i$ and intraindividual measurement errors $\mathbf{e}_i$ are normally distributed. However, Tong and Zhang (2012) claimed that both random effects and intraindividual measurement errors may be nonnormal.

### Two Types of Outlying Observations

Although there is no rigid mathematical definition of what constitutes an outlying observation, a commonly accepted characterization is that outlying observations are observations that do not follow the distributional pattern of the majority of data. The existence of outlying observations in GC modeling may due to extreme scores in either or both of $\mathbf{e}_i$ and $\mathbf{u}_i$. Because extreme scores in $\mathbf{e}_i$ or $\mathbf{u}_i$ affect the model estimation differently (Tong and Boker, 2016), it is necessary to distinguish different types of outlying observations in GC modeling. In factor analysis, Yuan and Zhong (2008) defined observations whose factor scores are far from the center

of the majority of the factor scores as leverage observations, and defined outliers as observations whose measurement errors are large, regardless of the values of the corresponding factor scores. They suggested that similar definitions can be used in other structural equation models. Following the definitions in Yuan and Zhong (2008), we distinguish two types of outlying observations in GC modeling. First, when an outlying observation is caused by extreme scores in random effects ($\mathbf{u}_i$), it is called *a leverage observation*. The intraindividual measurement errors for a leverage observation may be small or large. The observation corresponding to a small measurement error is called a good leverage observation. It is called a bad leverage observation when the measurement error is large. Second, when an outlying observation is caused by extreme scores in intraindividual measurement errors ($\mathbf{e}_i$), it is called *an outlier*. Note that it is possible that there might be individuals with unusual values in both their measurement errors and growth coefficients. These individuals are both a leverage observation and an outlier.

To further illustrate the pattern differences among outlying observation caused by nonnormal random effects $\mathbf{u}_i$ and/or nonnormal measurement errors $\mathbf{e}_i$, we generate and plot data from four types of distributional models (see Figure 1). For each type of distributional model, data on 20 individuals are generated at four equally spaced time points with a linear growth trend. Figure 1(1) displays the trajectories of the data generated without any leverage observations or outliers. The overall trend looks clean and smooth. Figure 1(2) plots the data generated with outliers (i.e., intraindividual measurement errors contain extreme scores). Noticeably, some observations stand out of the overall trajectory such as those labeled by $a$ and $b$. A close look at the two observations reveals that they deviate from the overall trajectory because they are off their own expected growth trajectories. For example, an individual might perform unexpectedly well in a test because s/he cheated, but his/her overall rate of change was not substantially different from other individuals'. Figure 1(3) portrays data generated with leverage observations (i.e., random effects contain extreme scores). Some observations also deviate from the overall growth trajectory. However, those observations are still on their own expected growth trajectories. The reason why they stand out is that the rate of growth for the specific individual is very different

from others'. An example could be that some talented individuals may learn faster than the others. Figure 1(4) draws the trajectories for data generated with observations being both leverage observations and outliers (i.e., both intraindividual measurement errors and random effects contain extreme scores simultaneously). Obviously, the observations which stand out are due to two sources - the trajectory of an individual deviates from the overall trajectory and the observation for this specific individual is off its own expected trajectory. For example, the observation $e$ stands out because it is off the expected trajectory of the case and the case itself has a higher initial level.

As clearly shown in Figure 1, leverage observations and outliers lead to different patterns of growth trajectories. This emphasizes again why it is important to distinguish the two types of outlying observations in GC modeling . In sum, leverage observations are caused by extreme scores in $\mathbf{u}_i$ and outliers are caused by extreme scores in $\mathbf{e}_i$, and in general, we call leverage observations and outliers together as outlying observations in GC modeling. We would like to note that in this article, **we use the term "outlier" when only measurement errors in GC models have extreme scores, and the term "outlying observation" is more general and used whenever an observation has extreme scores**.

Insert Figure 1 here

Diagnostics of outlying observations in GC modeling are very important in order to obtain a better model estimation. It is equally important and maybe more meaningful to identify leverage observations and outliers. For example, Tong and Boker (2016) claimed that some robust methods may perform well when data contain outliers, but they should be used more carefully when data contain leverage observations. In addition, leverage observation detection can be used to identify talented students whose growth trajectories are different from the average trajectory, and outlier detection can be used to detect test fraud, a very serious and popular practical task. If a student

took a series of tests in a period of time and got preternatural scores in one or two tests, s/he might be a suspected cheater. Since these topics are important in social, behavioral and educational researches, we apply methods to distinguish the two types of outlying observations in our study.

## Outlying Observations Versus Influential Observations

Observations may also be examined for influential status. Influential observations are defined by their impact on parameter estimates or/and the overall model fit. In contrast, an outlying observation is observed to be distributionally aberrant when comparing with other observations and is considered as being contaminated or coming from a different population. It has been demonstrated that an influential observation may not necessarily be an outlying observation, and vice versa. Therefore, the ideas of how to detect influential observations and outlying observations are different. A commonly applied method to detect influential observations is to delete the suspected observations and see how results are affected either at the level of overall model fit or at the level of parameter estimates. Whereas for methods used to detect outlying observations, a Mahalanobis distance is calculated from each point to the "center" of the data and an outlying observation is a point with a large distance.

The motivation of detecting influential observations and outlying observations is mainly to check whether there are observations that may potentially influence the model estimation and then determine some strategies to deal with these observations if necessary. Studies on influential case detection have been conducted in multilevel models where case deletion diagnostics were applied (e.g., Shi and Chen 2008; Van der Meer et al. 2006). Pek and MacCallum (2011) suggested to use multiple measures of case influence because cases may influence different aspects of results, and cases that exert little or no influence on one aspect may show a strong influence on another aspect. Another issue with case deletion is that it is affected by sample size (Pek and MacCallum, 2011). A large sample size leads to a high computation burden because $N$ ($N$ =total sample size) sets of model results need to be computed from $N$ delete-one-case samples, with each set of results then compared with results obtained from the full sample. More importantly, some observations may

have a joint effect. Multiple observations may have an influence on model fit or estimates of key parameters simultaneously, but deleting one of them each time does not change the model estimation, especially when sample size is large. Namely, sample size moderates the degree of influence that observations may exert on results. The joint effects of multiple observations can be taken care of by deleting the suspected multiple observations altogether, however, it is not feasible in practice as we never know which observations are influential observations before a detection method is applied, and it is extremely time consuming if we exhaustively try to test all combinations of observations. A forward search algorithm has been developed (Mavridis and Moustaki, 2008) and can release the computational burden, but it actually used the features of outlying observations. Therefore, detecting outlying observations is more practical as only observations that distributed differently need to be identified. Although using a measure such as Mahalanobis distance to screen for and delete outlying observations may not be effective and leave some highly influential observations in the sample (Pek and MacCallum, 2011), after leverage observations and outliers are defined distinctively, this problem can be largely resolved. This is because we assume that the model is correct and distinguishing leverage observations and outliers and detecting them separately can better find observations that deviate from the model. The effect of leverage observations and outliers on the parameter estimates and model fit in structural equation modeling is well understood. In particular, outliers can make the parameter estimates inconsistent, whereas good leverage observations have no effect on the likelihood ratio statistic but mainly affect the estimates of factor variances-covariances and the accuracy of factor loading estimates (Yuan and Zhong, 2008). Good leverage observations are influential to some fit indices such as CFI, NFI, and SRMR, but not to some other indices such as RMSEA, GFI, and adjusted GFI. Outliers and bad leverage observations are influential to all fit indices following NML (Yuan and Zhong, 2013). By identifying outliers and leverage observations correctly, highly influential observations are taken into account so that the masking effects can be greatly reduced.

## Six Methods for Outlying Observation Detection in GC Modeling

The detection of outlying observations in multivariate data is recognized to be an important but also difficult problem. Multivariate outlying observations usually exist when multiple measurements are obtained. Various methods can be used to detect outlying observations. Some are graphical such as normal probability plots. Others are model-based. In this section, six methods are proposed to identify multivariate outlying observations that deviate from the postulated GC model, among which two methods are GC model independent and the other four are GC model dependent. We successively discuss these methods below.

### GC Model Independent Methods

**1. Univariate detection (UD).** To detect multivariate outlying observations in a longitudinal dataset using the univariate detection method, we detect univariate outlying observations at each measurement occasion. Any case with univariate outlying observation(s) at one or more measurement occasions is considered as a multivariate outlying observation in GC modeling.

Several methods can be used to detect univariate outlying observations, among which the method based on interquartile range is commonly used. Let $Q_1$ and $Q_3$ be the lower and upper quartiles of a sample, respectively. One could define outlying observations to be the ones outside the range $[Q_1 - k(Q_3 - Q_1),\ Q_3 + k(Q_3 - Q_1]$ for some nonnegative constant $k$. The popular boxplot (or box-and-whisker plot) is based on this method with $k = 1.5$. We use this method to identify univariate outlying observation in this article.

The advantages of UD are obvious: the algorithm is easy to implement and the calculation is very fast. However, it also has disadvantages. Most importantly, because the procedure of univariate detection is as if we eyeball the observations and pick those with extreme scores at each measurement occasion, high dimensional outlying observations can be well hidden. A multivariate outlying observation can distort not only measures of location and scale but also those of correlation. Thus, with three or more dimensions, outlying observations can be difficult

or impossible to identify from coordinate plots of observed data. A simulated example is provided below for illustration. Two artificial datasets are generated. Dataset 1 (D1), including observations for 100 individuals at 4 time points, is generated from a traditional linear growth curve model with normal assumptions. The average latent slope $\beta_S$ of the overall trajectory is positive. Dataset 2 (D2) is generated by randomly replacing observations for 10 individuals in D1 with multivariate outlying observations. In particular, the observations for these 10 individuals are generated from a distinct linear growth curve model with slightly larger average latent intercept, negative average latent slope, and larger intraindividual measurement errors. The trajectory plots and boxplots of D1 and D2 are displayed in Figure 2. The trajectories for the 10 multivariate outlying observations in D2 are marked in red. Eyeball examination on those plots at each measurement occasion fails to locate suspected outlying observations, indicating that univariate detection methods are unable to detect multivariate outlying observations. In other word, UD is subject to masking effects. We fit a linear growth curve model to the two datasets, conduct NML estimation, and compare the average latent slope $\beta_S$ estimates. For D1, the average latent slope estimate is significantly different from 0, while for D2, it is not significant, indicating that unidentified multivariate outlying observations may lead to misleading statistical inferences.

Insert Figure 2 here

**2. Multivariate detection based on robust squared Mahalanobis distances (SMD).** Since UD may fail to identify multivariate outlying observations in many cases, multivariate detection methods have been developed. A univariate outlying observation may typically be thought of as the one that lies an abnormal distance from other values in a sample. The idea for multivariate detection is the same. We calculate a distance from each point to the "center" of the data. An outlying observation is a point with an extremely large distance. The distance is conventionally measured by squared Mahalanobis distance (M-distance), which is defined as

$$d^2(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), \tag{3}$$

where $\mathbf{x}_i$ is a $p$-dimensional observation for the $i$th individual ($i = 1, \ldots, N$ with $N$ representing the sample size), $\boldsymbol{\mu}$ is the population mean vector and $\boldsymbol{\Sigma}$ is the population covariance matrix. When data are multivariate normally distributed, squared M-distances follow a chi-square distribution with degrees of freedom $p$ (Mardia et al., 1979). Because the population mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are unknown in reality, they have to be estimated in order to get estimated squared M-distances by replacing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with their estimates, and the estimated squared M-distances approximate a chi-square distribution. Obviously, the sample mean and sample covariance matrix are not good estimates when outlying observations exist. Instead, robust estimators which are more resistant to outlying observations should be used. Among a variety of robust estimation methods that have been developed, the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985) is most widely used. Geometrically, covariance matrix specifies an ellipsoid that covers most data. Outlying observations stretch the ellipsoid toward the direction where the outlying observations are. MCD method is to find smaller volume of the ellipsoid to cover the majority data. Although other methods, such as finite sample reweighted-MCD and iterated reweighted-MCD, have been proved to outperform MCD under some circumstances (Cerioli, 2010), MCD estimator is still a respected and the most well known procedure for the following reasons. First, it asymptotically follows a normal distribution. Second, it is affine equivariant, so that measurement scale changes or other linear transformations do not alter the behavior of analysis methods. Third, MCD can be used easily because of the availability of a fast and efficient algorithm called FAST-MCD (Rousseeuw and van Driessen, 1999). Fourth, MCD method is built in statistical software such as R and SAS, so that it is convenient to use in practice.

By replacing the population mean and covariance matrix by the MCD estimates of them, the estimated squared M-distances are obtained. Outlying observations can then be identified by comparing the empirical distribution of squared M-distances with the corresponding chi-square distribution (e.g., Filzmoser et al., 2005; Rousseeuw, 1985). Several approaches can be implemented. For example, Garrett (1989) introduced the chi-square plot, which draws the

empirical distribution of the squared M-distances against the $\chi_p^2$ distribution. A break in the tail of the distributions is an indication for outlying observations, and values beyond this break are deleted so that a straight line appears. Rousseeuw and van Zomeren (1990) used a certain quantile (e.g., the 97.5% quantile) as a cutoff value for distinguishing outlying observations from non-outlying observations. Filzmoser et al. (2005) developed a method, which can be seen as an automation of Garrett (1989), by measuring the deviation of the data distribution from multivariate normal distribution in the tails. These approaches have been compared in Filzmoser (2005). Because the performances of them are comparable and are largely determined by the performance of the MCD estimator, we use the approach in Rousseeuw and van Zomeren (1990) in this article, as it is the easiest to understand and compute. The cutoff quantile is pre-determinted by us. If the quantile is high, the detection is more conservative. Otherwise if the quantile is low, the detection is more liberal. We use 97.5% quantile in this article. In practice, applied researchers may control how liberal the method is by adjusting the cutoff quantile.

Note that the GC model independent methods (i.e., UD and SMD), no matter taking into account of high dimensional outlying observations or not, cannot distinguish leverage observations and outliers of GC models.

## GC Model Dependent Methods

**3. Mean shift testing (MST).** Mean shift models and variance inflation models are regarded as two types of outlying-observation-generating models. The mean shift model is typically used to identify outlying observations to make them available for further study. The variance inflation model is often adopted for robust techniques with the aim of tolerating or accommodating outlying observations. Because the purpose of our study is to detect outlying observations, mean shift models are adopted. In practice, mean shift models are very commonly used (e.g., Barnett and Lewis, 1984; Rocke and Woodruff, 1996), so the problem of outlying observation detection can be reduced to testing whether or not the mean of the population is actually shifted if the suspected outlying observations are deleted from the original sample.

Therefore, the idea of MST is similar to case deletion diagnostics which are often used in influential observation detection. MST is developed by Pan and Fang (2002). The test is based on the generalized Cook's statistic, as Cook's distance provides an overall measurement of the change in all parameter estimates or a selection thereof (Cook, 1977). Let $D_i$ represent the generalized Cook's statistic (Pan and Fang, 2002, pp. 176-177) for the $i$th individual, $i = 1, \ldots, N$,

$$D_i = \left(\frac{Np_{ii}}{1-p_{ii}}\right)\left(\frac{\mathbf{r}_i'\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\mathbf{S}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\mathbf{r}_i}{1-p_{ii}}\right),$$

where $\boldsymbol{\Lambda}$ is the factor loading matrix as defined in Equation (1), $\mathbf{S} = \mathbf{Y}(\mathbf{I} - \mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z})\mathbf{Y}'$, $\mathbf{r}_i$ is the $i$th column of $\mathbf{Y}(\mathbf{I} - \mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z})$, and $p_{ii}$ is the $i$th diagonal element of the projection matrix $\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}$. The $1 \times N$ matrix $\mathbf{Z}$ consists of all ones for the typical GC model, that is $\mathbf{Z} = \mathbf{1}_{1\times N}$, and $\mathbf{Y}_{T\times N} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. Outlying observations can be identified by comparing the empirical distribution of $c_iD_i$ to a Beta distribution as

$$c_iD_i \sim Beta(\frac{N-q-1}{2}, \frac{q}{2}), \tag{4}$$

where $c_i = \dfrac{1-p_{ii}}{Np_{ii}}$ is a scalar specified for the $i$th individual. A certain quantile (e.g., the 97.5% quantile) of the Beta distribution can be used as a cutoff value. For individual $i$, if the calculated $c_iD_i$ is greater than the cutoff value of the Beta distribution, this individual is considered as an outlying observation of the GC model. Again the cutoff quantile is determined by applied researchers and controls how liberal the method is.

Similar to UD and SMD, MST still cannot distinguish leverage observations and outliers of GC models, because the mean shift could be due to extreme values either in intraindividual measurement errors or in the random effects, or in both.

**4. Multivariate detection based on individual-level growth curve analysis (IGC).** As pointed out by Bollen and Arminger (1991), observations are outlying observations because they are not well-predicted by the model, and individual-level residuals from latent variable models are one means to identify outlying cases. Following this idea, we propose to identify multivariate outlying observations in growth curve analysis based on individual-level growth coefficients and

residuals, and denote this method as IGC. In IGC, individual-level growth curve analyses ($\mathbf{y}_i = \mathbf{\Lambda} \cdot \mathbf{b}_i + \mathbf{e}_i,\ i = 1, \ldots, N$) are conducted. Namely, a regression model is fitted for each individual separately. Using least squares or maximum likelihood estimation methods, the individual coefficients $\mathbf{b}_i = (b_{i0}, \ldots, b_{iq})'$ are estimated and retained, denoted by $\hat{\mathbf{b}}_i$, and the residuals $\hat{\mathbf{e}}_i = (\hat{e}_{i1}, \ldots, \hat{e}_{iT})' = \mathbf{y}_i - \mathbf{\Lambda} \cdot \hat{\mathbf{b}}_i$ can be obtained accordingly. Let $\mathbf{B} = (\hat{\mathbf{b}}_1, \cdots, \hat{\mathbf{b}}_N)'$ and $\mathbf{E} = (\hat{\mathbf{e}}_1, \cdots, \hat{\mathbf{e}}_N)'$, so $\mathbf{B}$ is a $N \times q$ matrix of estimated individual coefficients and $\mathbf{E}$ is a $N \times T$ matrix of residuals for all individuals. Then, we would like to figure out which cases in $\hat{\mathbf{b}}_1, \cdots, \hat{\mathbf{b}}_N$ distributed differently from the rest cases and these cases are leverage observations. We also want to identify extreme cases in $\hat{\mathbf{e}}_1, \cdots, \hat{\mathbf{e}}_N$ and these cases are outliers. To achieve these goals, robust estimates of the mean vector and covariance matrix of $\mathbf{B}$ can be obtained through MCD method, based on which each individual's squared M-distance for individual coefficients $\hat{\mathbf{b}}_i$ is calculated. Individuals with extremely large squared M-distances for individual coefficients are leverage observations. Meanwhile, for each individual, we can also calculate robust squared M-distances for residuals based on the MCD estimates of the mean and covariance matrix of $\mathbf{E}$. Individuals with extremely large squared M-distances for residuals are outliers.

Notice that because of the collinearity of residuals, the covariance matrix of residuals is not of full rank and thus cannot be inversed to get squared M-distances. The residual-based squared M-distances has to be defined in a different way. Yuan and Zhong (2008) proposed that, for the covariance matrix of residuals, one get its eigenvectors corresponding to its zero eigenvalues. Then, one can find a matrix $\mathbf{A}$ whose columns are orthogonal to those eigenvectors. The covariance matrix of $\mathbf{A}\hat{\mathbf{e}}_i$ is of full rank. So, in IGC, residual-based squared M-distances are actually the squared M-distances for $\mathbf{A}\hat{\mathbf{e}}_i$.

**5. Non-robust model-based latent factor and residual analysis (NFRA).** Instead of fitting an individual-level growth curve model person by person, we also propose to fit one growth curve model to all data and use the individual-level random coefficients and residuals to detect outlying observations. This methods is denoted as NFRA. In the first step of this method, we fit a GC model to data and estimate the model by NML. Through Bartlett method (Bartlett, 1937),

factor scores (random coefficients) of the model can be obtained by

$$\hat{\mathbf{b}}_i = (\mathbf{\Lambda}'\hat{\mathbf{\Psi}}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\hat{\mathbf{\Psi}}^{-1}\mathbf{y}_i, \tag{5}$$

where $\hat{\Psi}$ is the estimated covariance matrix of $\mathbf{e}_i$. Based on $\hat{\mathbf{b}}_i$, the individual-level residuals can be easily calculated by subtracting $\mathbf{\Lambda}\hat{\mathbf{b}}_i$ from $\mathbf{y}_i$. Then, we can calculate robust squared M-distances for factor scores and individual-level residuals, and then compare the empirical distributions of them with chi-square distributions, separately, to find outlying observations in factor scores and individual-level residuals. The outlying observations for factor scores are leverage observations, while the outlying observations for individual-level residuals are outliers. Note that the covariance matrix of individual-level residuals is again not of full rank, so one needs to compute the M-distance in a sub-space as described in the previous section for IGC. Here, the Bartlett method is used because substituting Bartlett estimates for the latent factors does not lead to biased analysis when data are normally distributed (Yuan and Hayashi, 2010).

**6. Robust model-based latent factor and residual analysis (RFRA).** RFRA is similar to NFRA as discussed above, in which factor scores (random coefficients) and individual-level residuals are studied. However, in RFRA, factor scores and individual-level residuals are obtained through robust model estimation methods where potential outlying observations are downweighted with Huber-type weights (Yuan and Zhang, 2012b). Although it seems logically paradoxical to use robust methods to detect outlying observations as the influence of outlying observations is reduced in robust analysis, it is actually reasonable because by downweighting potential outlying observations, the estimated means and covariance matrices are closer to the population means and population covariance matrix. Therefore, the calculated factor scores and individual-level residuals are more like those in the population. In this case, leverage observations and outliers can be detected more precisely.

In RFRA, individual-level residuals are obtained by a direct robust method, while factor scores are obtained by a two-stage robust method in order to minimize the effects of both leverage observations and outliers (see more details in Yuan and Zhong, 2008). Moreover, the squared M-distances of factor scores and residuals for each individual are calculated differently in RFRA

from in NFRA. In NFRA, they are estimated with MCD estimators of the mean vectors and covariance matrices, whereas in RFRA, they are obtained by directly using the estimated means and covariance matrices of the factor scores and residuals from the robust methods.

Note that Methods 4-6 can distinguish leverage observations and outliers. Besides, Methods 5 and 6 can be easily generalized to outlying observation detection for other structural equation models. We would also like to make it explicit that MCD estimators are used in methods SMD, IGC, NFRA and RFRA. The only difference is that MCD estimator makes use of raw data in SMD method, whereas in the other three methods, it makes use of individual coefficients and measurement errors.

## Performance Evaluation of the Six Methods through a Simulation Study

We have discussed six methods to detect multivariate outlying observations in GC modeling. The goal of this study is to systematically evaluate and compare the performance of them. It is achieved through a Monte Carlo simulation study, by focusing on a linear unconditional GC model, which is a special case of the general GC model where

$$\mathbf{\Lambda}=\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T \end{pmatrix},\ \mathbf{b}_i = \begin{pmatrix} b_{Li} \\ b_{Si} \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} \beta_L \\ \beta_S \end{pmatrix},\ cov(\mathbf{e}_i) = diag(\sigma^2_{e_1}, \ldots, \sigma^2_{e_T}),\ \text{and}\ cov(\mathbf{u}_i) = \begin{pmatrix} \sigma^2_L & \sigma_{LS} \\ \sigma_{LS} & \sigma^2_S \end{pmatrix}.$$

The subscripts *L* and *S* refer to the initial level and slope, respectively. Note that the diagnostic methods can be easily extended to curvilinear and other nonlinear functional forms of the GC models or conditional GC models with time-varying and/or time-invariant covariates.

In the linear GC model, the population parameter values are selected as a subset of those in Tong and Zhang (2012) and are given below.

$$\beta_L = 6,\ \beta_S = 2,\ \sigma^2_L = 1,\ \sigma^2_S = 1,\ \sigma_{LS} = 0,\ \sigma^2_{e_j} = 1,\ j = 1, \ldots, T.$$

We conducted pilot studies and found that different values of $\sigma^2_L$, $\sigma^2_S$, $\sigma_{LS}$, and $\sigma^2_{e_j}$ do not affect the performance of the six diagnostic methods. So, we fix the values of those parameters in this Monte Carlo simulation study.

## Design Conditions

The probability of detecting outlying observations depends on many factors. For example, Rocke and Woodruff (1996) concluded that problems are more difficult when sample size is small, the proportion of outlying observations is large, or outlying observations are concentrated. Moreover, when data dimension is high, the MCD estimator used to estimate robust M-distances may break down. Based on the previous literature, in this study, five potentially influential factors are manipulated including sample size, number of measurement occasions (i.e., dimension), proportion of outlying observations, geometry of outlying observations, and type of outlying observations. The sample size is 50, 100, 300, 500, or 1000, ranging from a small sample size to a large one. We expect that with larger sample size, outlying observations are easier to identify so the sensitivities of the detection methods would be higher. The number of measurement occasions is 4, 5, or 8. When more measurement occasions are included, the data dimension is higher so the MCD estimator might break down. We would like to investigate whether MCD estimator is still effective when the number of measurement occasions is as high as 8. Conditions for other three factors are discussed below in the explanation of data generation process.

Given a certain sample size and a number of measurement occasion, we generate data from the unconditional linear GC model with normal assumptions as given previously. This dataset, denoted as O0, does not contain any outlying observation and is retained as a comparison to the other conditions. We use mean shift models to generate outlying observations in this study for three reasons. First, the data generated from the mean shift models often distributed differently from the original data and follow the definition of outlying observations. Second, as mentioned previously, mean shift models are regarded as one of the most common outlying-observation-generating models, in which the outlying values are generated from a distribution with the same covariance matrix and a shifted mean. For example, in a longitudinal study, some individuals may have higher initial levels or faster rates of change than the majority of the individuals. These individuals' scores can be viewed as from a GC model with a relatively larger $\beta$ but same covariance matrices as the rest of individuals. Third, shift outlying observations

provide a reasonable test bed for multivariate outlying observation detection (Rocke and Woodruff, 1996). To generate outlying observations, we randomly select 2%, 5%, or 10% individuals from the dataset O0. The percentage is the proportion of outlying observations and we replace the observations for these individuals by outlying values. For the geometry of outlying observations, we consider generating outlying values from a normal distribution with its mean 2, 4, or 6 standard deviations away from the center of the majority of the data. It is hypothesized that the farther the outlying observation is away from the center of the majority of the data, the easier it can be identified by the proposed methods. For each dataset, it may contain one of three types of outlying observations: (1) both leverage observations and outliers, (2) outliers only, or (3) leverage observations only, and the dataset is denoted as O1, O2, or O3, accordingly. Basically, after a certain proportion (2%, 5%, or 10%) of individuals are randomly selected in O0, we generate O1, O2, and O3 by substituting these individuals' scores in different ways. For O1, we equally divide the randomly selected individuals into three groups. We re-generate individuals' scores in group 1 from a mean shift model with the means of both $\mathbf{e}_i$ and $\mathbf{u}_i$ being shifted. We re-generate individuals' scores in group 2 from a mean shift model with only the mean of $\mathbf{e}_i$ being shifted and re-generate individuals' scores in group 3 from a mean shift model with only the mean of $\mathbf{u}_i$ being shifted. For O2, observations for all the random selected individuals are re-generated from a mean shift model with the mean of $\mathbf{e}_i$ being shifted. For O3, the selected individuals' observations are re-generated from a model with only the mean of $\mathbf{u}_i$ being shifted. Note that for $\mathbf{e}_i$, the mean shift can occur at only one measurement occasion, or more measurement occasions, and for $\mathbf{u}_i$, the mean shift can be at the latent intercept, or the latent slope, or both. The six diagnostic methods will be used to detect outlying observations and distinguish leverage observations and outliers. We expect that UD performs the worst.

In summary, we have 5 conditions for sample size, 3 conditions for number of measurement occasions, 3 conditions for outlying observation proportion, 3 conditions for outlying observation geometry, and 3 conditions for outlying observation type. Overall, 420 conditions of simulations are investigated. For each condition, we evaluate the six diagnostic methods based on 500

replications. [1]

## Evaluation Criteria

Sensitivity and specificity are the statistical measures that we use to evaluate the performance of the six diagnostic methods. Sensitivity (also called the true positive rate) measures the proportion of positives that are correctly identified as such. Specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified as such. In our study, sensitivity measures how likely an outlying observation can be identified as an outlying observation, while specificity measures the probability of a non-outlying observation being correctly identified as a non-outlying observation.

$$Sensitivity = \frac{number\ of\ true\ positives}{total\ number\ of\ outlying\ observations}$$

$$Specificity = \frac{number\ of\ true\ negatives}{total\ number\ of\ non-outlying\ observations}$$

The closer the sensitivity and the specificity are to 1, the better the diagnostic method is. For a statistical test, sensitivity is essentially statistical power and specificity is $1 - Type\ I\ error\ rate$. Therefore, the nominal specificity should be the cutoff quantile for methods SMD and MST in detecting outlying observations. For methods IGC, NFRA, and RFRA, the nominal specificities in detecting leverage observations and outliers are the cutoff quantiles. When an outlying observation is mistakenly identified as a good observation, we say that there is a masking effect. When good data are mistakenly identified as outlying observations, there is a swamping effect. Thus, masking problems exist when sensitivity is low, while swamping problems need to be considered when specificity is low.

For the six diagnostic methods, we can compare their sensitivities and specificities in detecting outlying observations. Moreover, for IGC, NFRA, and RFRA, we further compare their sensitivities and specificities in detecting leverage observations and outliers.

[1] A pilot study was conducted with 1000 replications. The results are the same.

## Results

According to our simulation results, the number of measurement occasions is not a significant factor when the proportion of outlying observations is as low as those being set in this study. Although sensitivity and specificity for each method are slightly smaller for $T = 8$ than those for $T = 4$ and $T = 5$, the different values in $T$ do not cause notable differences in the performance of the six diagnostic methods. Therefore, the following results are presented regarding the other four factors when the number of measurement occasions $T = 4$. The results for $T = 5$ and $T = 8$ are available upon request. We first compare the six methods in detecting all the outlying observations. Then, we focus on the last three methods, IGC, NFRA, and RFRA, comparing their performance in detecting outliers and leverage observations, separately.

**Outlying observation identification in general.** Table 1 presents specificities of the six methods in detecting outlying observations in O0, which is the dataset without any outlying observation, under different sample sizes. Note that for O0, sensitivity is unavailable to measure. With the increase of sample size, specificities for the six methods are getting closer to 1. The fact that specificities are not exactly 1 could due to the methods themselves or sampling errors. Among the six methods, UD and MST perform slightly worse as they have more severe swamping problems by identifying more non-outlying observations as outlying observations. When sample size is small (e.g., 50 or 100), SMD and NFRA have much lower specificities than the other four methods, meaning that they are sensitive to sample size and are not suggested to use with small samples. By comparing the results from NFRA and RFRA, we suggest using RFRA instead of NFRA as robust methods provide more reliable detection results. From the table, it seems that IGC and RFRA always perform better and may be trusted.

Insert Table 1 here

For datasets containing outlying observations, both sensitivity and specificity of each

method are calculated under every condition. Figures 3 and 4 present sensitivities and specificities of the six methods in detecting outlying observations in O1 (datasets containing both leverage observations and outliers), O2 (datasets only containing outliers), and O3 (datasets only containing leverage observations), respectively, when the proportion of outlying observations is 2% and 10%, and the outlying values are generated with the mean 2 and 6 standard deviations away from the center of the majority of the data. [2] Each figure is organized to have 2 rows and 6 columns, and consists of 12 subfigures. From the top row to the bottom row, the proportion of outlying observations is increased from 2% to 10%. Columns 1 and 2, columns 3 and 4, and columns 5 and 6 can be viewed as three separate blocks and the three blocks display the outlying observation detection results for O1, O2, and O3, respectively. From the left to the right within each block, the mean shift of the outlying observation generating model is increased from 2 to 6. In each subfigure, sensitivities or specificities of the six diagnostic methods are displayed at different sample sizes. The vertical dotted lines in light grey shows the sample sizes we consider in this study. We evaluate the effects of the four factors - sample size, proportion of outlying observations, geometry of outlying observations, and type of outlying observations, on the performance of the six diagnostic methods. First, sample size does not substantially influence sensitivities of the six methods. By looking at Figures 3, we notice that the six lines which represent the six diagnostic methods in each subfigure are almost flat, meaning that a larger sample size does not lead to a larger sensitivity value for each method and will not reduce the problem of masking. However, larger sample size can reduce the problem of swamping, since specificities of most methods increase along with the increase of sample size. In Figures 4, we see steep upward climbs of the lines for the detection methods, especially when sample sizes are small. Second, by comparing the rows of each figure, it seems that the proportion of outlying observations is not very influential to the performance of the six methods. Although it is true that sensitivities and specificities of those methods are slightly better when the proportion of outlying observations is lower, the differences are hardly noticeable. Note that if the proportion is higher

[2]The complete simulation results for all study conditions are available upon request.

than $1/(T+1)$, it would have a greater influence and the six diagnostic methods may break down. Third, the geometry of outlying observations has a great effect on sensitivities of the six methods, but almost no effect on specificities. If the outlying observation comes from a distribution whose mean is far away from the majority of the data, it is easy to identify. Otherwise, if the outlying observation comes from a distribution which overlaps a lot with the distribution for non-outlying observations, it may not be able to be detected. For example, when the mean shift of the outlying observation generating model is 2 standard deviations away from the center of the majority of the data, sensitivities of the methods are around 0.2 or below under all conditions, indicating that all six diagnostic methods have problems of masking and should not be trusted. Fourth, the type of outlying observations also influence sensitivities of the six diagnostic methods substantially. By comparing the three blocks in Figure 3, we conclude that leverage observations are much easier to identify than outliers as sensitivities in the second block are about twice or even more times bigger than those in the third block for some detection methods. Even when the mean shift of the outlying observation generating model is 4 standard deviations away from the mean of the majority of the data, sensitivities of the some methods can still be as low as 0.2 in detecting outliers.

Next, we take a closer look at the figures and compare the performance of the six diagnostic methods. It is obvious that UD has lower sensitivity and specificity under most conditions, meaning that univariate method is not suggested to detect multivariate outlying observations. SMD and NFRA perform similarly. Both are liberal and have higher sensitivity but lower specificity, indicating that they are good at recognizing outlying observations, but they may also mistakenly treat non-outlying observations as outlying observations and cause swamping problems. Moreover, the specificities of them are greatly influenced by sample size. When sample size is small, both methods have more severe swamping problems. MST is very conservative as its sensitivity is the lowest among all six methods, especially in detecting outliers. Although the specificity of MST is higher than that for the other methods, the difference is subtle. IGC has high specificities, especially when sample size is large. It also has higher sensitivities

among the six methods when the outlying observation is far away from the center of majority of the data. However, if the distribution of the outlying observation is close to the distribution of most data, IGC can perform worse than the other methods in recognizing outlying values. RFRA is comparable to IGC, with reasonably high sensitivities and specificities. Another advantage of RFRA is that its performance does not seem to be related to sample size. The detection results from RFRA are more stable for small samples.

Insert Figure 3 here

Insert Figure 4 here

**Outlier identification.** IGC, NFRA, and RFRA can distinguish outliers and leverage observations. Thus, we investigate the performance of them in detecting outliers first and detecting leverage observations next. Figures 5 presents sensitivities and specificities of the three methods in detecting outliers for O1, O2, and O3, when the proportion of outlying observations is 5%. The results for 2% and 10% are similar and thus omitted for the sake of saving space. For O3, the datasets do not contain any outliers, so sensitivities are unavailable to measure. This is why the right block in Figure 5 only consists of specificities. As shown in left and middle blocks of the figure, NFRA has a high sensitivity in detecting outliers, meaning that it is good at picking outliers out from the datasets. Since this method is liberal, it has a relatively lower specificity, and may lead to swamping problems. Comparing IGC and RFRA, we find that RFRA almost always has a higher sensitivity, and their specificities are about the same. In addition, RFRA is more stable for small sample sizes. Therefore, RFRA overall performs better than IGC.

Insert Figure 5 here

**Leverage observation identification.** Figures 6 presents sensitivities and specificities of the three methods in detecting leverage observations for O1, O2, and O3, when the proportion of outlying observations is 5%. The results for 2% and 10% are similar and thus omitted to save space. For O2, the datasets do not contain any leverage observation, so sensitivities are unavailable. Thus, the middle block in Figure 6 only consists of specificities. It seems that RFRA performs better in identifying leverage observations as its sensitivity is higher under almost all conditions and its specificity is about the same as the specificities for other two methods. Moreover, IGC and NFRA have low specificities when sample size is small, while RFRA is more stable to small sample sizes. So, RFRA is also more reliable in detecting leverage observations.

Insert Figure 6 here

## An Example

In this section, we illustrate the application of the six outlying observation diagnostic methods through analyses on a subset of data from the National Longitudinal Survey of Youth 1997 (NLSY97) Cohort (Bureau of Labor Statistics, U.S. Department of Labor, 2005). The dataset contains 512 school children's Peabody Individual Achievement Test (PIAT) mathematics scores yearly from the 7th grade to the 10th grade. The individuals' trajectory plot (Figure 7) suggests a linear growth pattern for the development of math abilities. The boxplot (Figure 8) indicates potential outlying observations and the PIAT math scores at each year are skewed to the left. Results from both D'Agostino skewness test (D'Agostino, 1970) and Anscombe-Glynn kurtosis test (Anscombe and Glynn, 1983) show that the skewness and kurtosis at each measurement occasion are significantly different from those of normal distributions. Because the

data are nonnormal and may contain potential outlying observations, we use this dataset to illustrate the application of outlying observation detection methods.

Insert Figure 7 here

Insert Figure 8 here

The six diagnostic methods are applied, and the outlying observations detected by them are given in Table 2. To facilitate the application of the detection methods by applied researchers, we provide corresponding R codes in the Appendix. Outlying observations detected by UD are most different from those detected by all the other methods. Among the 26 identified outlying observations, 7 of them (1, 2, 10, 30, 36, 509, and 512) were not detected as outlying observations by the other methods. We may infer that the specificity for UD is low. SMD and NFRA identify most outlying observation: 8.2% individuals are outlying observations, and the results from SMD and NFRA are identical. In addition, NFRA detect 2 individuals as both leverage observations and outliers. MST detect fewest outlying observations. This is consistent with our simulation results as the sensitivity for MST is always the lowest. The results from IGC and RFRA are close to each other, but RFRA detects more leverage observations. According to the simulation results, because RFRA has higher sensitivity in detecting leverage observations, we should trust the results from RFRA as more reliable. Thus, among the 512 school children, 7 of them are leverage observations and have growth patterns different from the majority of the cases; and 22 of them are outliers with extreme values of intraindividual measurement errors (as shown in Figure 9). We may delete or downweight those outlying observations before conducting the data analysis or directly use robust methods to avoid biased parameter estimates and misleading statistical

inferences. More discussion on how to use multiple methods to correctly identify outlying observations will be provided in the discussion section.

Insert Table 2 here

Insert Figure 9 here

## Discussion

Six outlying observation diagnostic methods in growth curve modeling are evaluated in this article, including two GC model independent methods (UD and SMD) and four GC model dependent methods (MST, IGC, NFRA, and RFRA). Among these methods, IGC, NFRA, and RFRA can be used to distinguish outliers and leverage observations, where outliers represents extreme values at the intraindividual measurement errors and leverage observations represents extreme values at the random effects (i.e., latent coefficients). A Monte Carlo simulation study is conducted, by manipulating five potentially influential factors, including sample size (50, 100, 300, 500, and 1000), number of measurement occasions (4 and 8), proportion of outlying observations (2%, 5%, and 10%), geometry of outlying observations (mean shift can be 2, 4, or 6), and type of outlying observations (leverage observation, outlier, or both). Among these factors, the number of measurement occasions and the proportion of outlying observations do not substantially influence the performance of the six diagnostic methods under the studied simulation conditions. The following conclusions can be drawn for the other three factors. First, sample size does not have a big effects on the sensitivities of the six methods, although increasing sample size can greatly improve specificities of some methods, such as SMD and NFRA. Second, the geometry of outlying observations is an important factor to detect outlying observations. If the

outlying values are far away from the center of the majority of the data, they are more likely to be identified. Third, leverage observations are easier to detect than outliers, especially when outlying values are close to good data.

According to our simulation results, UD is not recommended to use as it has lower sensitivity and specificity under most conditions. SMD usually has high sensitivities and can detect the most number of outlying observations. However, it may lead to swamping problems as good data can also be identified as outlying values. In addition, it is sensitive to small samples. MST has a low rate to detect outlying values. Theoretically, an advantage of MST is that it can test whether a set of individuals are outlying observations or not simultaneously. So, if a certain set of individuals are suspected to have extreme values, we may use MST to test their scores all at once. However, just as the case deletion diagnostics for influential observation detection, this is not realistic in practice because we don't know potential outlying observations prior to conducting the diagnostic methods and it is impossible to test all combinations of individuals. So, we calculate all individuals' generalized Cook's statistics and compare them to the cutoff value. In fact, Pan and Fang (2002) suggested a different way in conducting MST. In the first step, generalized Cook's statistics for all individuals are calculated, and the largest one could be determined. If this value is less than the critical value of the nominal Beta distribution, one concludes that there is no outlying observation in the data set. Otherwise, the corresponding individual is an outlying observation. One deletes that individual and repeats the above process for the remaining data. If the largest Cook's statistic is again above a cutoff value, one take a look at this individuals' scores together with those just been deleted and test whether they are outlying observations as a whole. The algorithm stops when there is no longer any outlying observation in the remaining data. We compared this approach to the approach discussed previously in the method section through simulation and found that they provide similar results. Therefore, we only present one approach in the main part of this article because the presented approach is simpler and computationally faster. For the three methods that can be used to distinguish outliers and leverage observations, NFRA is liberal and can identify most outlying values, but it may also lead to

swamping problems as good data are incorrectly identified as outlying observations. In addition, NFRA is sensitive to small samples. Although RFRA and IGC behave similarly, RFRA usually performs better in a small degree, with a slightly higher sensitivity in most situations. It is worth mentioning that RFRA is more robust to small samples, so the results from RFRA should be weighted more if sample size is small. In addition, note that the above conclusions are drawn by assuming that the model is true. When models are misspecified, the model independent methods (UD and SMD) still perform the same. However, the performance of the model dependent methods may be affected and their performance for misspecified models should be further studied.

The mean shift model was used to generate outlying observations in this study since Rocke and Woodruff (1996) suggested that the hardest kind of outlying observations to find is the kind that has a covariance matrix with the same shape as the good data. Although pure shift outlying observations might seem to be detectable, they usually cannot be identified by eyeball examination and in fact, no method is known that can find the outlying observations with complete assurance. It is always true that outlying observation diagnostic methods have problems of masking and swamping. Basically, they may overlook some outlying values, or mistakenly recognize some good data as outlying observations. If there is masking problems, the dataset still contain outlying observations and thus the nonnormality still cause inconsistent and inefficient parameter estimates. Swamping seems to be an acceptable side effect in some situations, however, there are applications where even a moderate amount of swamping may have disastrous consequences (see Cerioli, 2010 for more detailed examples). Therefore, we should be cautious about both masking and swamping. The greatest chance of success comes from use of multiple methods. Like what we did in the real data example, we compare the results from all the detection methods except UD, take a close look at those observations on which the five methods provide different diagnostic conclusions, and then make a careful decision based on our experiences and the purpose of the study. If our purpose is to obtain unbiased parameter estimates, it is better to be more liberal and detect as many outlying observations as possible. However, if we want to retain a high statistical power, or detect some abnormal behaviors or ethical issues, the swamping problem

should be avoided.

We would like to note that the robust MCD estimators are used to estimate squared M-distances. Although MCD has been proved to outperform minimum volume ellipsoid estimator in Woodruff and Rocke (1994), there are other estimators such as reweighted MCD that has been shown to perform better. Moreover, since the rejection rule to detect outlying observations often leads to an inflated Type II error, Hardin and Rocke (2005) developed more precise cutoff values to improve the performance of MCD estimators in detecting multivariate outlying observations. They proposed that the estimated squared M-distance approximates an F distribution better than a chi-square distribution for small sample sizes, even when data are multivariate normal. In our study, the MCD estimator was chosen because it is most frequently used and available in standard statistical software packages. But the six diagnostic methods discussed in this article can also base on other estimators for population mean vector and covariance matrix. When the dimension of data increases, the bias of the MCD estimates grows almost exponentially. In this case, a high-breakdown method (e.g., Cerioli, 2010) which can deal with a substantial fraction of outlying observations in the data should be resorted to. In addition, it is known that the estimates of random coefficients and intraindividual measurement errors may have shrinkage in GC modeling (e.g., Morris and Lysy, 2012). Parameters that are estimated with small accuracy shrink more than very accurately estimated parameters. In this article, several diagnostic methods for outlying observation detection perform very well in the simulation for the unconditional linear GC model even without considering the shrinkage. When the model is more complicated, shrinkage might affect the performance of outlying observation identification. In those cases, some techniques such as computing a range of plausible values may build in sampling variability to avoid shrinkage.

We also want to point out that the cutoffs used to determine whether a data point is an outlying observation are fixed at 97.5th percentiles of the corresponding Chi-square distributions in the simulation study. By selecting different cutoffs, there is a tradeoff between sensitivities and specificities. How to find out the optimal cutoffs can be further investigated in the future. In

addition, although multiple outlying observations are detected simultaneously by the proposed methods, the simultaneity adjustments when comparing multiple distances to the relevant cutoff value is absence in this article. Previous literature (e.g., Becker and Gather, 2001) suggested Bonferroni-type adjustments of the asymptotic chi-square distribution of the robust M-distances, however, these corrections will cause low powers. Other studies (e.g., Pan and Fang (2002) as described above) have suggested to test multiple outlying observations in steps. Based on our simulation results, it is time consuming and provides similar results as our current approaches.

After the outlying observations are identified, different strategies can be applied to deal with them. Popular techniques that have been suggested include deleting outlying observations, downweighting outlying observations, data transformation, and robust methods. If the nonnormality of data is caused by some nonnormal distribution, data transformation and robust methods may perform better in handling such data. If the nonnormality is due to data contamination or outlying observations, deletion or downweighting techniques as well as some robust methods may perform well. Because in practice it is never known whether the nonnormality is a result of a nonnormal distribution or data contamination, robust methods are recommended to use under many circumstances. In addition, if the proportion of detected outlying observations in the data is large, a mixture model may be more recommended to apply. We would like to further point out that Tong and Boker (2016) recently showed that if an outlying observation is a leverage observation in GC modeling, deletion technique performs better than some robust methods. Note that this statement is based on the assumption that the extreme values in random coefficients (i.e., a leverage observation) in GC modeling are not a property belonging to the population. For example, researchers who study the effect of a training program probably do not want to treat talented students as a part of the population. In such a case, deleting those talented students from the data may provide a more reasonable interpretation of the training effect than using the robust method does. However, if an outlying observation is an outlier, those robust methods provide fairly good model estimation results. Therefore, it is important to distinguish outliers and leverage observations as different strategies need to be adopted to handle them. This

article provides ways to identify and distinguish outliers and leverage observations in GC modeling.

To summarize, this article systematically studied six outlying observation diagnostic methods in growth curve modeling. The univariate detection method is not suggested to use when multivariate outlying observations exist. We recommend to use multiple methods among the other five multivariate detection methods, compare their results, and make a decision based on research questions. We also emphasize the importance to distinguish leverage observations and outliers. Among the three methods which can detect leverage observations and outliers, RFRA is more reliable. Furthermore, both NFRA and RFRA can be easily extended to outlying observation diagnosis for general structural equation models.

## Appendix

R codes for the real data example:

```
##univariate outying observation detection function
uniout <- function(data){
        F.l<-quantile(data, .25)
        F.u<-quantile(data, .75)
        d.F<-F.u-F.l
        C.l<-F.l-d.F*1.5
        C.u<-F.u+d.F*1.5
        res <- c(which(data<C.l),which(data>C.u))
        res
}

##multivariate outying observation detection function for the SMD method
mdout <- function(data,alpha=0.05){
```

```
        mu <- cov.rob(data,method="mcd")$center
        sig <- cov.rob(data,method="mcd")$cov
        md2 <- diag(t(t(data)-mu)%*%solve(sig)%*%(t(data)-mu))
        cut <- qchisq((1-alpha/2),4)
        mdo <- as.numeric(which(md2>cut))
        mdo
}


##read the dataset into R
y      <-      read.table('nlsy.txt')

T <- ncol(y)
N <- nrow(y)

## method 1: UD

m1 <- sort(c(uniout(y[,1]),uniout(y[,2]),uniout(y[,3]),uniout(y[,4])))
m1.o <- as.numeric(unique(m1))
dput(m1.o)      #outlying observations

## method 2: SMD

m2.o <- mdout(y)
dput(m2.o)      #outlying observations

#method 3: MST
```

```
m <- 2
r <- 1

z <- t(rep(1,N))
pz <- t(z)%*%solve(z%*%t(z))%*%z
S <- t(y)%*%(diag(N)-pz)%*%y
E <- t(y)%*%(diag(N)-pz)
M <- lambda%*%solve(t(lambda)%*%S%*%lambda)%*%t(lambda)

Tvec <- rep(NA,N)

for(i in 1:N){
        pii <- pz[i,i]
        ei <- E[,i]
        Tvec[i] <- (t(ei)%*%M%*%ei)/(1-pii)
}
TT <- sort(Tvec, decreasing=TRUE)
Tindex <- order(Tvec, decreasing=TRUE)

cf <- qf(.975,m,N-r-m)
cv <- m*cf/(N-r-m+m*cf)

m3.o <- Tindex[which(TT>=cv)]

m3.o <- sort(m6.o)
dput(m3.o)      #outlying observations
```

```
## method 4: IGC

lambda <- cbind(rep(1,T),0:(T-1))
res <- matrix(NA,N,T)
b <- matrix(NA,N,2)
for(i in 1:N){
        b[i, ] <- solve(t(lambda)%*%lambda)%*%t(lambda)%*%y[i,]
        res[i,] <- y[i,]-lambda%*%b[i,]
}

ind <- which(eigen(cov(res))$values<1e-6)
eigvec <- eigen(cov(res))$vectors[,ind]
A <- semdiag.orthog(eigvec)
nres <- t(A)%*%t(res)

m4.o <- mdout(t(nres))
m4.l <- mdout(b)
dput(m4.o)      ##outliers
dput(m4.l)      ##leverage observations


#method 5: NFRA
library(lavaan)

colnames(y) <- c('y1','y2','y3','y4')
```

```
gcmodel<-'i =~ 1*y1 + 1*y2 + 1*y3  + 1*y4
                    s =~ 0*y1 + 1*y2 + 2*y3 + 3*y4'

res.lavaan <- growth(gcmodel, data=data.frame(y))
fs <- predict(res.lavaan)
ym <- x%*%t(fs)
resid <- y-t(ym)

m5.o <- mdout(resid)
m5.l <- mdout(fs)
dput(m5.o)        ##outliers
dput(m5.l)        ##leverage observations

detach(package:lavaan)

#method 6: RFRA
library(semdiag)
lgcm<-specifyModel()
          b0 -> y1, NA, 1
          b0 -> y2, NA, 1
          b0 -> y3, NA, 1
          b0 -> y4, NA, 1
          b1 -> y1, NA, 0
          b1 -> y2, NA, 1
          b1 -> y3, NA, 2
     b1 -> y4, NA, 3
          b0 <-> b0, sb0, NA
```

```
        b1 <-> b1, sb1, NA
        b0 <-> b1, sb01, NA
        y1 <-> y1, s1, NA
        y2 <-> y2, s2, NA
        y3 <-> y3, s3, NA
        y4 <-> y4, s4, NA

yout.1<-try(semdiag(y, ram.path=lgcm, max_it = 10000,software='sem'))
out <- semdiag.summary(yout.1)
m6.o <- as.numeric(c(out[[3]],out[[1]]))
m6.l <- as.numeric(c(out[[2]],out[[1]]))
dput(m6.o)      ##outliers
dput(m6.l)      ##leverage observations
```

## References

Anscombe, F. J. and Glynn, W. J. (1983). Distribution of kurtosis statistic for normal statistics. *Biometrika*, 70:227–234. DOI: 10.2307/2335960.

Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data*. New York : John Wiley and Sons.

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology*, 28:97–104. DOI: 10.1111/j.2044–8295.1937.tb00863.x.

Becker, C. and Gather, U. (2001). The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Computational Statistics and Data Analysis*, 36:119–127. dio: 10.1016/S0167–9473(00)00032–3.

Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software, Encino, CA.

Bollen, K. A. (1987). Outlier and improper solutions: a confirmatory factor analysis example. *Sociological Methods and Research*, 15:375–384. DOI: 10.1177/0049124187015004002.

Bollen, K. A. and Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methods and Research*, 21:235–262. DOI: 10.2307/270937.

Bureau of Labor Statistics, U.S. Department of Labor (2005). *National Longitudinal Survey of Youth 1997 cohort, 1997-2003 (rounds 1-7) [computer file]*. OSU, Produced by the National Opinion Research Center, the University of Chicago and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, Ohio.

Cadigan, N. G. (1995). Local influence in structural equation models. *Structural Equation Modeling*, 2:13–30. DOI: 10.1080/10705519509539992.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, 105:147–156. DOI: 10.1198/jasa.2009.tm09147.

Cook, R. (1977). Detection of influential observations in linear regression. *Technometrics*, 19:15–18. DOI: 10.2307/1268249.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society: Series B*, 48:133–169.

D'Agostino, R. B. (1970). Transformation to normality of null distribution of g1. *Biometrika*, 57:679–681. DOI: 10.2307/2334794.

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. *Austrian Journal of Statistics*, 34:127–138. DOI: 10.17713/ajs.v34i2.406.

Filzmoser, P., Garrett, R. G., and Hauser, H. (2005). Multivariate outlier detection in exploration geochemistry. *Gomputers and Geosciences*, 31:579–587. DOI: 10.1016/j.cageo.2004.11.013.

Garrett, R. G. (1989). The chi-square plot: A tool for multivariate outlier recognition. *Journal of Geochemical Exploration*, 32:319–341. DOI: 10.1016/0375–6742(89)90071–X.

Hardin, J. and Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14:928–946. DOI: 10.1198/106186005X77685.

Lee, S.-Y. and Wang, S.-J. (1996). Sensitivity analysis of structural equation models. *Psychometrika*, 61:93–108. DOI: 10.1007/BF02296960.

Lieberman, E. S. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, 99:435–452. DOI: 10.1017/S0003055405051762.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. New York : Academic Press.

Mavridis, D. and Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behavioral Research*, 43:453–475. DOI: 10.1080/00273170802285909.

McArdle, J. J. (1998). Modeling longitudinal data by latent growth curve methods. In Marcoulides, G., editor, *Modern methods for business research*, pages 359–406. Lawrence Erlbaum Associates, Mahwah, NJ.

McArdle, J. J. and Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. American Psychological Association.

Meredith, W. and Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55:107–122. DOI: 10.1007/BF02294746.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1):156–166. DOI: 10.1037/0033–2909.105.1.156.

Morris, C. N. and Lysy, M. (2012). Shrinkage estimation in multilevel normal models. *Statistical Science*, 27:115–134. DOI: 10.1214/11–STS363.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55(2):463–469. DOI: 10.1111/j.0006–341X.1999.00463.x.

Osborne, J. W. and Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research and Evaluation*, 9:1–12.

Pan, J.-X. and Fang, K.-T. (2002). *Growth curve models and statistical diagnostics*. Springer, New York.

Peña, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation (with discussion). *Technometrics*, 43:286–310. DOI: 10.1198/004017001316975899.

Pek, J. and MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: cases and their influence. *Multivariate Behavioral Research*, 46:202–228. DOI: 10.1080/00273171.2011.561068.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91:1047–1061. DOI: 10.2307/2291724.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, Vol. B, edited by W. Grossmann, G. Pflug, I. Vincze, and W. Werty, Reidel, Dordrecht, Netherlands:283–297.

Rousseeuw, P. J. and van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223. DOI: 10.2307/1270566.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–651. DOI: 10.2307/2289995.

Savalei, V. and Falk, C. (2014). Robust two-stage approach outperforms robust fiml with incomplete nonnormal data. *Structural Equation Modeling*, 21:280–302. DOI: 10.1080/10705511.2014.882692.

Shi, L. and Chen, G. (2008). Case deletion diagnostics in multilevel models. *Journal of Multivariate Analysis*, 99:1860–1877. DOI: 10.1016/j.jmva.2008.01.023.

Tong, X. and Boker, S. M. (August, 2016). The impact of masking and swamping effects for multivariate outlier diagnosis in structural equation modeling. *The 2016 Annual Convention of the American Psychological Association (Paper presentation)*, Denver, CO.

Tong, X. and Zhang, Z. (2012). Diagnostics of robust growth curve modeling using student's t distribution. *Multivariate Behavioral Research*, 47:493–518. DOI: 10.1080/00273171.2012.692614.

Van der Meer, T., Te Grotenhuis, M., and Pelzer, B. (2006). Influential cases in multilevel modeling: a methodological comment. *American Sociological Review*, 75:173–178. DOI: 10.1177/0003122409359166.

Woodruff, D. L. and Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association*, 89:888–896. DOI: 10.2307/2290913.

Yuan, K.-H. and Bentler, P. M. (2000). Inferences on correlation coefficients in some classes of nonnormal distributions. *Journal of Multivariate Analysis*, 72:230–248. DOI: 10.1006/jmva.1999.1858.

Yuan, K.-H. and Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, 54:161–175. DOI: 10.1348/000711001159366.

Yuan, K.-H. and Hayashi, K. (2010). Fitting data to model: Structural equation modeling diagnosis using two scatter plots. *Psychological Methods*, 15:335–351. DOI: 10.1037/a0020140.

Yuan, K.-H. and Zhang, Z. (2012a). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika*, 77:803–826. DOI: 10.1007/s11336–012–9282–4.

Yuan, K.-H. and Zhang, Z. (2012b). Structural equation modeling diagnostics using r package semdiag and eqs. *Structural Equation Modeling*, 19:683–702. DOI: 10.1080/10705511.2012.713282.

Yuan, K.-H. and Zhong, X. (2008). Outliers, high-leverage observations and influential cases in factor analysis: Minimizing their effect using robust procedures. *Sociological Methodology*, 38:329–368. DOI: 10.1111/j.1467–9531.2008.00198.x.

Yuan, K.-H. and Zhong, X. (2013). Robustness of fit indices to outliers and leverage observations in structural equation modeling. *Psychological Methods*, 18:121–136. DOI: 10.1037/a0031604.

Zhang, Z., McArdle, J. J., and Nesselroade, J. R. (2012). Growth rate models: emphasizing

growth rate analysis through growth curve modeling. *Journal of Applied Statistics*, 39:1241–1262. DOI: 10.1080/02664763.2011.644528.

Table 1

*Specificities of the six diagnostic methods in detecting outlying observations in O0 (dataset without any outlying observation)*

| | 50 | 100 | 300 | 500 | 1000 |
|---|---|---|---|---|---|
| UD | 0.953 | 0.968 | 0.974 | 0.975 | 0.976 |
| SMD | 0.894 | 0.953 | 0.976 | 0.979 | 0.980 |
| MST | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 |
| IGC | 0.956 | 0.980 | 0.990 | 0.992 | 0.993 |
| NFRA | 0.886 | 0.952 | 0.975 | 0.979 | 0.980 |
| RFRA | 0.981 | 0.980 | 0.980 | 0.980 | 0.980 |

Table 2

*Identified outlying observations in PIAT math data through the six diagnostic methods. For IGC, NFRA, and RFRA, ID numbers followed by a star indicate leverage observations, while ID numbers without a star indicates outliers. If an ID number is in parentheses, the corresponding individual is detected as both a leverage observation and an outlier.*

| | Total # (%) | Outlying observation IDs |
|---|---|---|
| UD | 26 (5.08%) | 1, 2, 3, 4, 5, 6, 7, 9, 10, 15, 19, 22, 28, 30, 36, 55, 71, 87, 200, 202, 244, 455, 507, 509, 510, 512 |
| SMD | 42 (8.20%) | 3, 4, 6, 7, 9, 14, 15, 19, 22, 23, 26, 28, 40, 54, 55, 56, 71, 78, 87, 139, 161, 200, 202, 229, 244, 275, 295, 299, 345, 379, 395, 403, 441, 454, 455, 461, 471, 482, 484, 488, 507, 510 |
| MST | 10 (1.95%) | 3, 5, 6, 7, 19, 87, 229, 484, 488, 510 |
| IGC | 24 (4.69%) | 4, 6*, 7, 15, 28, 40, 56, 71, 78, 87*, 200, 202, 229*, 244, 295, 299, 345, 359, 379, 395, 403, 455, 461, 482 |
| NFRA | 42 (8.20%) | 3, 4, (6*), 7, 9, 14, 15, 19, 22, 23, 26, 28, 40, 54, 55, 56, 71, 78, 87, 139, 161, 200, 202, 229, 244, 275, 295, 299, 345, 379, 395, 403, 441, 454, 455, 461, 471, 482, 484, 488, 507, (510*) |
| RFRA | 29 (5.66%) | 4, 5*, 6*, 7, 15, 19*, 28, 40, 56, 78, 87*, 200, 202, 229*, 244, 295, 299, 345, 379, 395, 403, 413, 441, 454, 455, 461, 482, 488*, 510* |

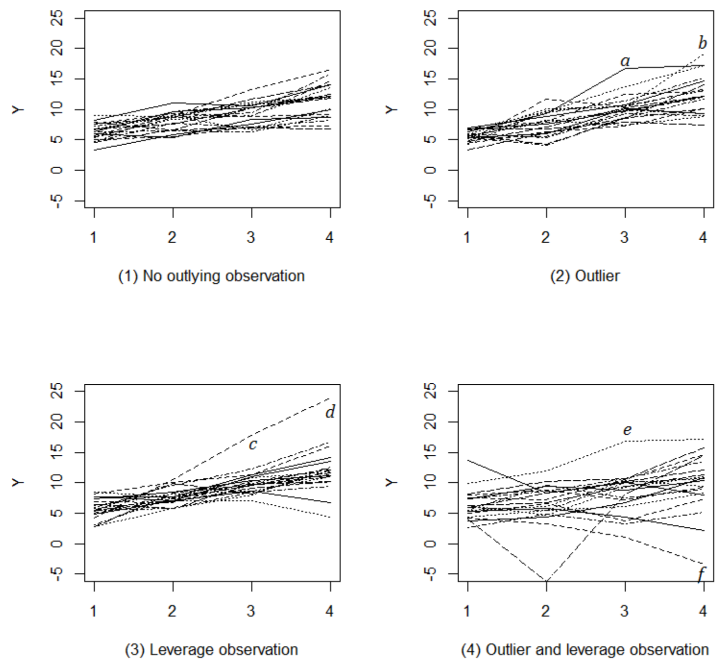


*Figure 1.* Trajectory plots of data generated without outlying observation, with only outliers, with only leverage observation, and with both. Data on 20 individuals are generated at 4 measurement occasions.

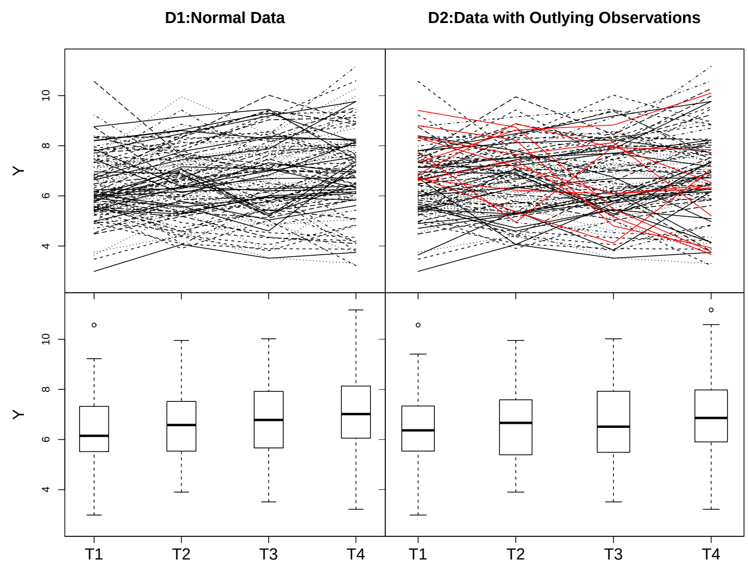


*Figure 2*. The trajectory plots and boxplots of two simulated datasets

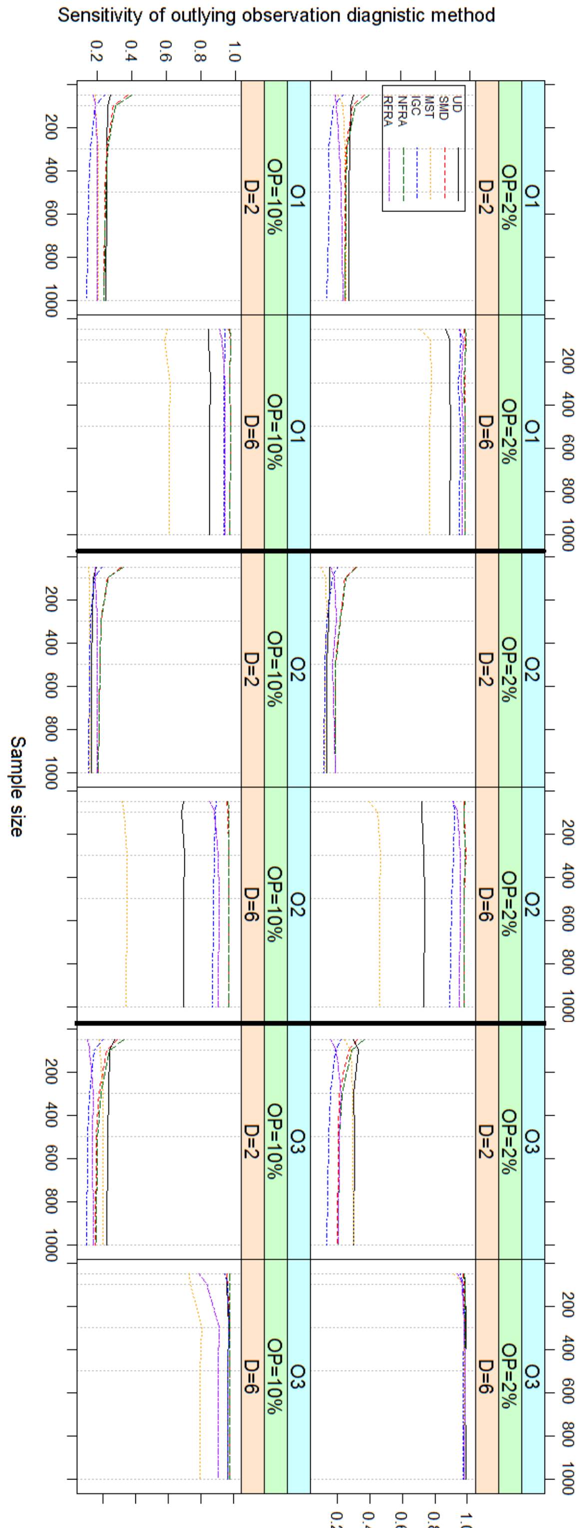

*Figure 3*. Sensitivities for outlying observation diagnostic methods for O1 (datasets containing both leverage observations and outliers), O2 (datasets only containing outliers), and O3 (datasets only containing leverage observations). OP denotes outlying observation proportion. D denotes the mean shift of the outlying observation generating model from the original model.

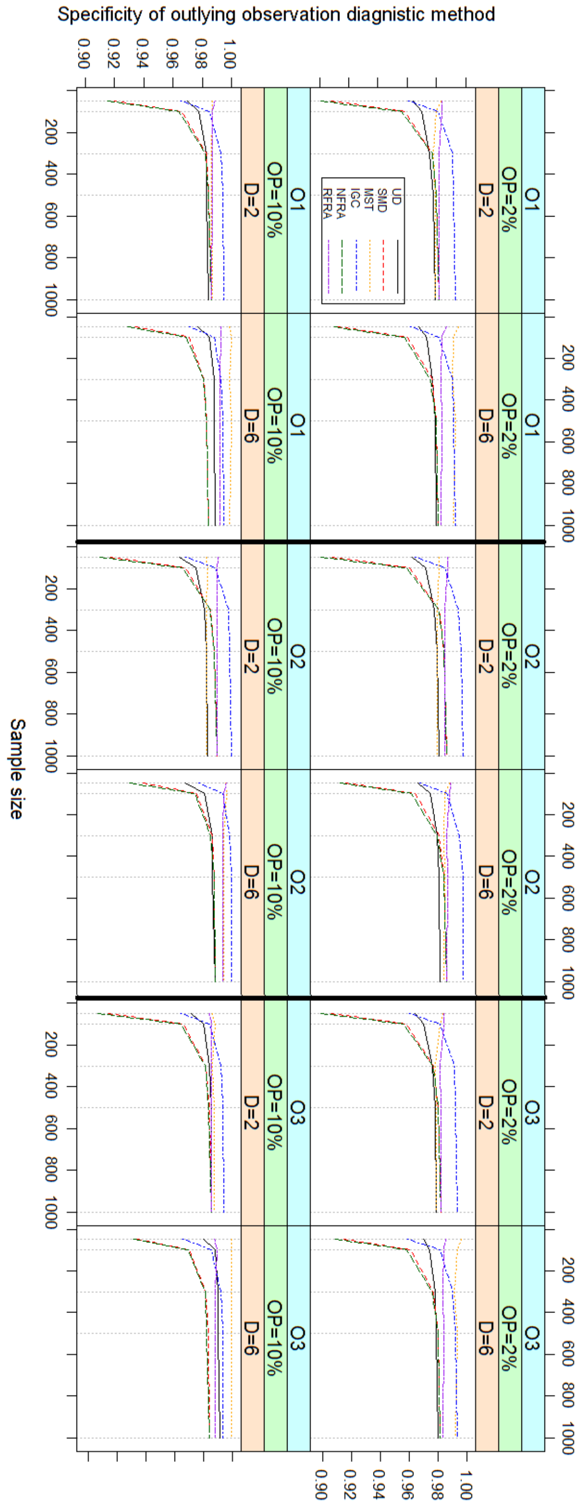


*Figure 4*. Specificities for outlying observation diagnostic methods for O1 (datasets containing both leverage observations and outliers), O2 (datasets only containing outliers), and O3 (datasets only containing leverage observations).

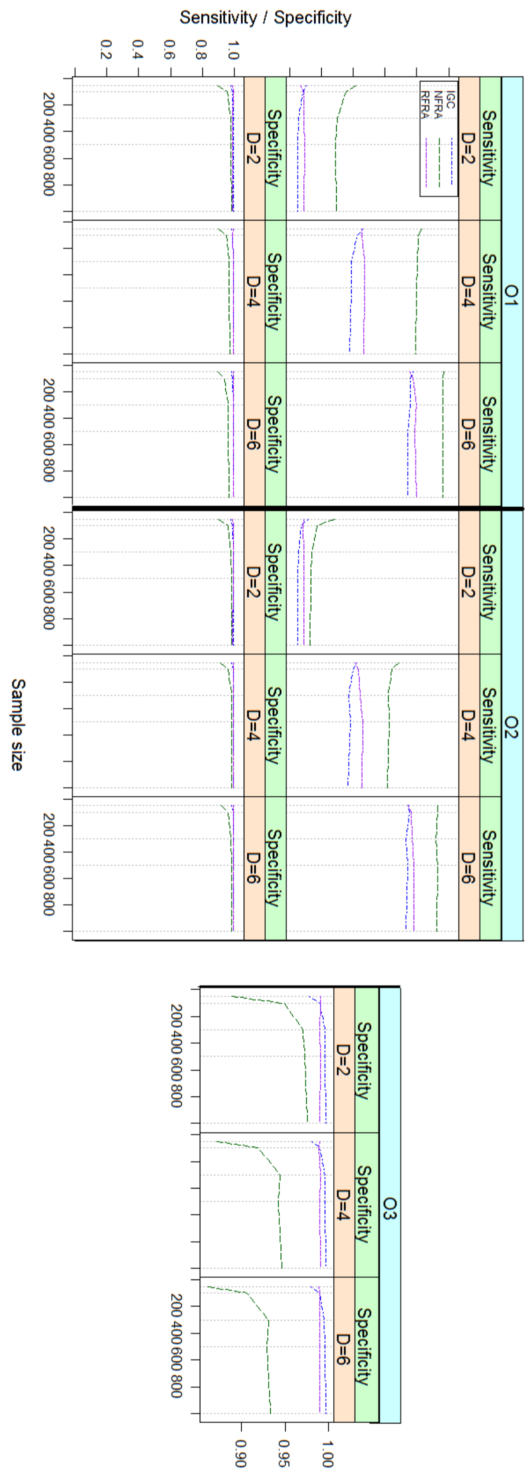


*Figure 5.* Sensitivities and specificities of IGC, NFRA, and RFRA in detecting outliers for O1 (datasets containing both leverage observations and outliers), O2 (datasets only containing outliers), and O3 (datasets only containing leverage observations), when the proportion of outlying observation is 5%.

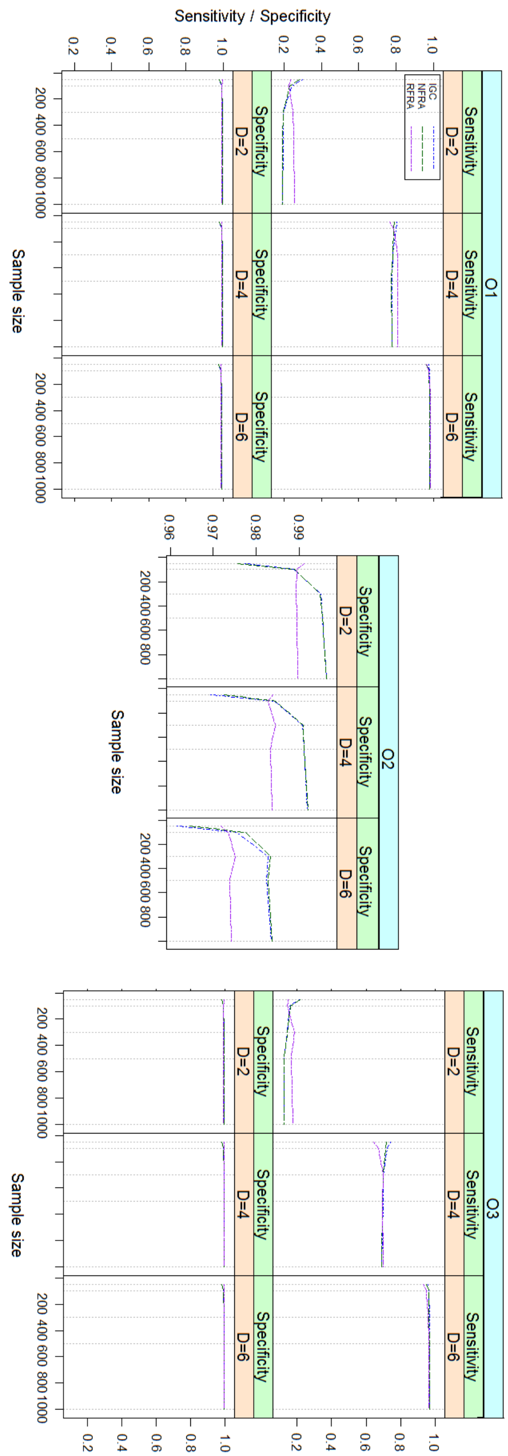


*Figure 6*. Sensitivities and specificities of IGC, NFRA, and RFRA in detecting leverage observations for O1 (datasets containing both leverage observations and outliers), O2 (datasets only containing outliers), and O3 (datasets only containing leverage observations), when the proportion of outlying observation is 5%.

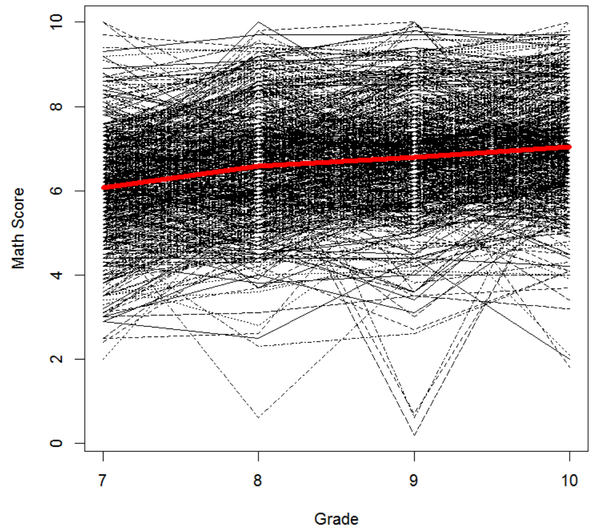


*Figure 7*. A collection of individual trajectories for the PIAT math data from NLSY97. 512 school children are measured at 4 occasions.

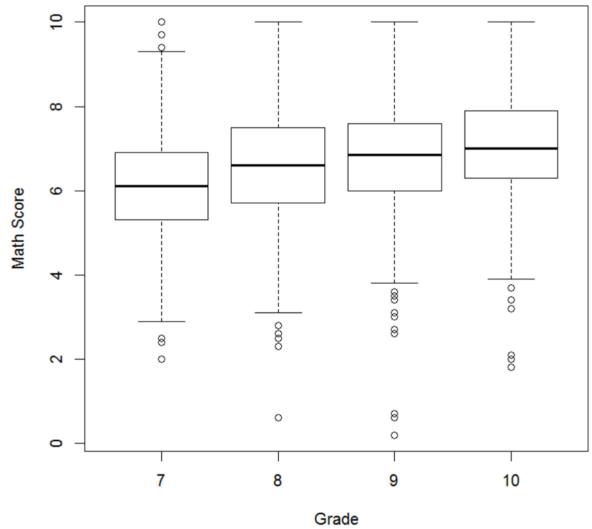


*Figure 8*. Boxplot for the PIAT math data from NLSY97. Circles represent potential outliers.

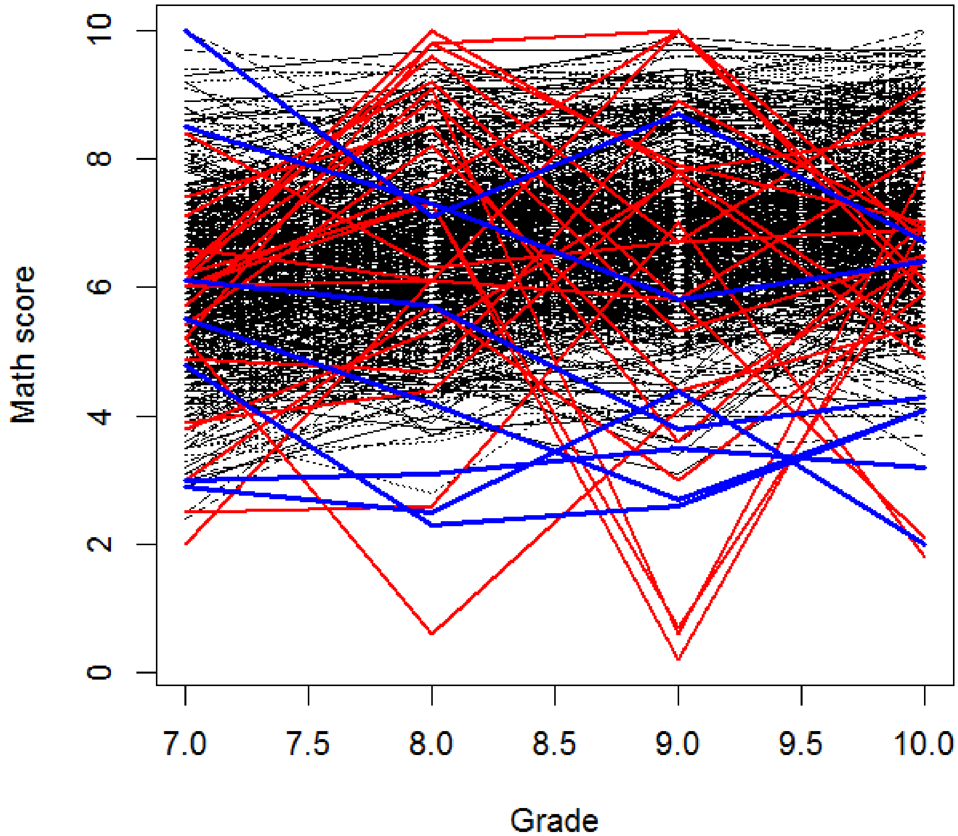


*Figure 9.* A collection of individual trajectories for the PIAT math data from NLSY97. Identified leverage observations are marked in blue and identified outliers are marked in red.